Journal of Documentation

Introduction to bibliometrics for construction and maintenance of thesauri: Methodical considerations

Jesper W. Schneider, Pia Borlund,

Introduction to bibliometrics for construction and maintenance of thesauri

Methodical considerations

Jesper W. Schneider and Pia Borlund
Royal School of Library and Information Science, Aalborg, Denmark

Received October 2003
Revised March 2004
Accepted May 2004

Keywords *Knowledge management, Controlled language construction, Cataloguing*

Abstract *The paper introduces bibliometrics to the research area of knowledge organization – more precisely in relation to construction and maintenance of thesauri. As such, the paper reviews related work that has been of inspiration for the assembly of a semi-automatic, bibliometric-based, approach for construction and maintenance. Similarly, the paper discusses the methodical considerations behind the approach. Eventually, the semi-automatic approach is used to verify the applicability of bibliometric methods as a supplement to construction and maintenance of thesauri. In the context of knowledge organization, the paper outlines two fundamental approaches to knowledge organization, that is, the manual intellectual approach and the automatic algorithmic approach. Bibliometric methods belong to the automatic algorithmic approach, though bibliometrics do have special characteristics that are substantially different from other methods within this approach.*

Introduction

Traditionally, knowledge organization within library and information science denotes classification, indexing, and cataloguing, applied to storage, access, and retrieval of documents[1] in information retrieval (IR) systems (Anderson and Pérez-Carballo, 2001a). According to Anderson and Pérez-Carballo (2001b) all classification and indexing is based on classing or clustering of items[2] based on similarities of characteristics. Anderson and Pérez-Carballo (2001a) further point out that the term "clustering" implies automatic processes, in contrast to the term "classing" that usually implies human judgement. Whether the focus is on actual indexing, classification, or the construction of organizing systems the main issue is the same, that is, the description of content in order to group items of similar characteristics (Anderson, 1997). Basically, there are two fundamental approaches to the description of document content and consequently to construction and maintenance of knowledge organization systems – manual intellectual analysis, and automatic algorithmic analysis (Anderson, 1997; Lancaster, 1998; Adair, 19555; 2001b). These two approaches are increasingly combined to benefit from the strengths of each approach, and to counterbalance their weaknesses as well (Anderson, 1997). Manual intellectual construction and maintenance of knowledge organization systems is recommended when dealing with languages due to the dynamic and complex nature of language (e.g. Blair, 1990; Aitchison *et al.*, 2000). However, the manual intellectual approach is a resource demanding and costly process, which motivates to do research on less resource demanding construction and maintenance methods (Aitchison *et al.*, 2000; Anderson and Pérez-Carballo, 2001a; 2001b). Motivated by this, the overall aim of the research project is to verify the

Journal of Documentation
Vol. 60 No. 5, 2004
pp. 524-549

0022-0418
DOI 10.1108/00220410410560609

applicability of using bibliometric methods as a supplement to intellectual manual based construction and maintenance of thesauri as a type of knowledge organization systems. In relation to this, the present paper is mainly a review of related work and serves as an understanding of the potential methods employed in the research process of verifying the applicability of bibliometrics within thesauri construction and maintenance. The presented methods are by no means exhaustive.

The paper is composed of six main sections. The current section is the introduction. The next section presents in brief the research project and subsequently the focus of the paper. The methodical considerations concern the range and identification of appropriate bibliometric methods to employ in the bibliometric study of the identification of candidate thesaurus terms and concept relationships.

The subsequent sections three and four present background information and related work in relation to the research project. Thus, section three defines the research area of bibliometrics and introduces central aspects, such as types of citation analyses and theories of citing. Where as section four describes related work of traditional approaches to construction of thesauri, as well as previous research that uses bibliometric methods for knowledge organization.

Section five discusses appropriate bibliometric methods for identification of candidate thesaurus terms, and suitable network analytical and multivariate statistical techniques intended for detection of concept relationships, that is, the methodical considerations. The case used for demonstration and discussion is the topical specialty area of periodontology.

The sixth and final section, the summary statement, acknowledges the characteristics of bibliometrics as an additional source of input for thesaurus construction within knowledge organization. Further, it stresses that the selected range of bibliometric methods, network analyses, and multivariate statistical techniques are to be seen as possible, but not the definitive range of methods and techniques. Furthermore, it points out that the presented methods and techniques are not limited to construction of thesauri only, but they may be applicable to all types of knowledge organization systems. Finally, the section emphasises the motivation and idea of the research project.

Presentation of the research project: "application of bibliometrics to thesaurus construction and maintenance"

The ambition of the research project is to introduce bibliometric methods to the research area of knowledge organization for the purpose of thesaurus construction. Bibliometrics is the study of documents and their bibliographic reference and citation structures (Egghe and Rousseau, 1990, p. 203). Bibliometric methods are introduced as the basis of a semi-automatic approach to thesaurus construction, and are to be seen as a supplement to the resource demanding approach of manual intellectual thesaurus construction. Bibliometric methods are considered valuable for two reasons. One reason is that bibliometric methods, in contrast to the traditional automatic algorithmic methods for thesaurus construction, do not initially depend on the frequency distribution of index terms, but on the frequency distribution of citations and bibliographic references in documents. Cited documents act as "concept symbols" (Small, 1978; Rees-Potter, 1989), thus, concepts and conceptual relations based on cited documents has an advantage over concepts and conceptual relations created from conventional co-term analysis. Obviously, the identification of concepts and their

relations based on citations and references is independent of language and changing terminology (Leydesdorff, 1997). Another reason is that bibliometric methods may point to common topical characteristics of documents and their authors, and may be used to uncover, otherwise hidden knowledge structures of a discipline and its users (e.g. Borgman, 1990).

Consequently, the underlying hypothesis of the research project is that bibliometric methods can be used as a supplement to the established methods of thesaurus construction, because the bibliometric methods may uncover conditions, patterns and relationships between documents and their concepts. Based on the hypothesis, the research project aims at investigating:

- the ability of bibliometric methods to help in construction and maintenance of a thesaurus vocabulary and structure;
- the extent to which these methods can identify synonyms and uncover relations between terms; and
- the use of bibliometrics for maintenance of a thesaurus in a given domain over a given time period.

The focus of the present paper is the background and rationale for a semi-automatic approach used to investigate the applicability of bibliometric methods for thesaurus construction and maintenance. The approach employs a range of bibliometric methods, network analyses, and multivariate statistical techniques for the identification of candidate thesaurus terms and concept relationships. However, these methods are used in close conjunction with manual construction work.

Due to the extensive application of bibliometrics in the present paper (and project) combined with the relatively rare application of bibliometrics to the area of knowledge organization, the following section introduces bibliometrics in general, and more specifically with reference to central aspects of citation analysis and underlying practice for, and theories of citing.

Introduction to bibliometrics

Bibliometrics is the study of quantitative aspects of production, dissemination, and use of recorded information (Tague-Sutcliffe, 1992, p. 1). Bibliometrics encompasses a number of empirical methods, such as bibliographic coupling and co-citation analysis (Kessler, 1963; Small, 1973). Today, the field of bibliometrics include all quantitative aspects and models of science communication, storage, dissemination, and retrieval of scientific information (Glänzel and Schoepflin, 1994). The last definition integrates all presently existing orientations, such as applications to science policy, library science, IR, and knowledge organization in its broad context. The idea of using bibliometric methods in connection with construction and maintenance of knowledge organization systems is not new, but only few actual construction attempts have been made, e.g. the work by Rees-Potter (1989) and Garfield (1990), 1994). Bibliometric methods have been successfully applied to examine the intellectual structure of several disciplines (Small, 1977; White and Griffth, 1981; Borgman, 1990; White and McCain, 1998).

A major component of bibliometrics is citation analysis. The practice of scientific citing, and subsequent citation analysis, is based on assumptions and norms. Thus, the following two sub-sections describe first types of citation analyses, and subsequently the theories of citing.

Citation analysis

A citation is a reference to a document given by a more recently published document. The document citing is the *citing document*, and the document that receives the citation is the *cited document* (Smith, 1981; King, 1987). Citation analysis involves counting the number of citations to a particular document for a period of time after its publication (this is sometimes called direct citations) (Smith, 1981; King, 1987). The traditional understanding of the citation function is that the frequency with which a document is cited can be taken as a measure of the impact or influence of that document on the citing literature (Garfield, 1979). Citation analysis leads to more sophisticated methods, such as co-citation analysis (Small, 1973), mapping of the literature (e.g. Small and Griffith, 1974; Small, 1977; White and Griffth, 1981; White and McCain, 1998), bibliographic coupling (Kessler, 1963), and co-word analysis (Callon *et al.*, 1983). These methods, individually and in combination, strides to find information patterns, by analysing reference and citation patterns as well as word use frequencies, combined with statistical analysis. According to Narin *et al.* (1994) bibliometric analyses have three axioms concerning measurement: activity measurement, impact measurement, and linkage measurement. However, citation analysis in general presents a number of serious theoretical and methodical problems, hence the focus on the existing debate of theories of citing.

Theories of citing

Garfield (1998) uses the term "citationology" for the study of theory and practice of citations and citation analysis. According to Leydesdorff (1998) a variety of contexts for citation analysis exist, but no comprehensive theory of citation itself has been formulated. A citation is a complex unit and the citation process is a very complex one (Luukkonen, 1997). It requires an understanding of the underlying norms of the process, the different functions of citations, the quality of citations made, and the motivations and reasons for citing in general, to do bibliometric analyses (Cronin, 1984; King, 1987; Liu, 1993; Leydesdorff, 1998; Leydesdorff and Wouters, 1999). It is vital to understand these premises when choosing units for analyses, methods and measures, and subsequently interpreting the results. The prerequisites for doing bibliometric analyses are an agreement on the communication function of the text units employed, and on the measures applied to them (Wilson, 1999). This agreement has generated a great deal of debate within the field of bibliometrics. Basically, the debaters represent two different viewpoints to the understanding of the communicative function of citations. The one viewpoint is represented by the normative theory, and the other is the social-constructivist view (Cronin, 1984; Wilson, 1999).

According to Wilson (1999, p. 126) the normative interpretation of citations is that "[a] document is cited in another document because it provides information relevant to the performance and presentation of the research, such as positioning the research problem in a broader context, describing the methods used, or providing supporting data and arguments". The normative theory views citations as a formal registration of the use of specific findings embedded in a document selected on the basis of scientific merit. Authors who cite are, in effect, conditioned to follow the norms of science in general, and the norms of citation practice in their chosen fields of research in particular (e.g. Merton, 1973; Garfield, 1979; Price, 1986; Cole, 1992).

Based on the assumption that all citations are equal and that the individual citing is not necessarily exhaustive, but only sufficient for the author's purpose, some conclusions can be established (Wilson, 1999). The more a document is cited from a subsequent body of literature, the more the document influences the reported research. Hence, received citations can be employed as indicators of scientific impact, influence, or "quality" (Van Raan, 1998). A declining citation rate over time can suggest that the document's content is increasingly less relevant, i.e. that the document is becoming "obsolete" (e.g. Burton and Kebler, 1960; Line, 1993; Száva-Kováts, 2002)[3]. Further, if two documents are jointly cited by another document, they jointly contribute to the content and impact of that research document, and are associated by their role in that research document. Accordingly, the more two documents are co-cited from a body of literature, the greater is the association of their content, in the opinion of the authors of that body of literature. This leads to the co-citation analysis and its application in literature mapping and visualization studies (e.g. Small, 1973; White and Griffth, 1981; White and McCain, 1998; Börner *et al.*, 2003).

Different criticisms and counter-criticisms as well, of the normative theory of citing appear from time to time. In opposition to the normative theory's understanding of the communication function of citations is the social-constructivist view (e.g. Latour, 1987). In this view science is seen as a social process in which citations are mainly rhetoric instruments in order to persuade other scientists by non-logical means, such as to gain political advantage, advance own interests, defend one's claims against attacks, and convince others (e.g. Latour, 1987). This view alters the perception of the communicative function of citations, in that there are several, and that the rhetorical ones (also called ceremonial) dominate over the rewarding ones (Wilson, 1999). The social-constructivist view results in a highly critical attitude toward the use of citations as indicators of scientific performance (e.g. MacRoberts and MacRoberts, 1989a, 1989b, 1996; Cozzens, 1989; Seglen, 1998).

The social-constructivist interpretation of the communication function of citations is at present being counter-attacked through a variety of validation studies of citation analysis (White, 1990; Garfield, 1997; Kostoff, 1998; Van Raan, 1998; Vinkler, 1998). Further, Small (1998) points out that neither the normative theory of citing nor the social-constructivist view are easy to test, and that they both ignore the symbolic function of citations, i.e. the author uses the cited work to symbolize a particular idea inherent in that work (Garfield, 1965; Small, 1978). The fact that none of the viewpoints acknowledge the symbolic function of citations leads to analyses on citation content and citation context, where deeper interpretation is required than just the assignment of one typical communication function (e.g. Small and Greenlee, 1980; Small, 1982; Peritz, 1992; Liu, 1993).

In addition to the disputes over the communicative functions of citations, criticisms are also directed towards possible errors in citation formatting, content, and problems with the actual measures employed (King, 1987; Seglen, 1998; Wilson, 1999). Perfect compliance with the traditional normative view is not essential (Wilson, 1999; Borgman and Furner, 2002). The key issues are the degree to which the central claim is correct or incorrect, and how to validate/refute the competing interpretations. According to Wilson (1999, p. 130) "…the alternative methods (peer review, subject expert opinion etc.) do not have the status to unambiguously validate or repudiate citation analysis techniques".

In this section we have outlined assumptions and problem areas in relation to citation analysis and theories of citing, as it is vital to understand and comprehend these aspects and their possible influences when doing bibliometric studies.

Related work

The presentation of related work is to be seen in the context of the research project and focus of the paper, that is, construction and maintenance of thesauri by use of bibliometrics. Thus, we start with an introduction to the manual intellectual approach to thesaurus construction – defining the concept of a thesaurus. Subsequently, we present characteristics of and pointers to the research on automatic algorithmic thesaurus construction. In continuation of the sub-section of research on automatic algorithmic thesaurus construction, we close the section of related work with examples of bibliometrics employed to knowledge organization.

Manual intellectual thesaurus construction and maintenance

A thesaurus is a controlled indexing vocabulary that formally displays a priori relationships between concepts (Aitchison *et al.*, 2000). Traditionally, a thesaurus functions as an indexing and/or retrieval tool, helping with the selection of terms. Thesauri might differ in detail but they share a basic principle, in that they record a set of terms (word or phrases) covering some knowledge domain, and three types of relationships – equivalence, hierarchical and associative relationships (Miller, 1997).

Thesaurus construction requires collecting a set of terms (preferable nouns and noun phrases), and subsequently terminological and semantic treatment as well as relational structuring of the collected term classes (Soergel, 1974; Aitchison *et al.*, 2000). The classes cover restricted topics of specific scope, and collectively they cover the complete subject area in question. Thus, thesauri are fundamentally linguistic and conceptual in nature (Miller, 1997). Structural, semantic and terminological problems are ever present, and manual intellectual construction work is necessary when dealing with these problems (Aitchison *et al.*, 2000). This is usually done by scanning the subject literature for candidate terms and/or, by a group of experts who review the subject matter, suggest potential terms and propose reasonable class arrangements (Lancaster, 1998). A major disadvantage inherent to the use of any thesaurus, due to the dynamic nature of language, is the necessity to maintain the thesaurus. The approach taken to meet the requirements of thesaurus construction and maintenance in a less resource demanding way is the automatic algorithmic approach.

Automatic algorithmic thesaurus construction and maintenance

Automatically generated thesauri contain classes that reflect the usage of words in text corpora. Two basic corpus methods exist to automatic thesaurus construction, statistical co-occurrence analysis, and linguistic context analysis. The most traditional method in IR is statistical co-occurrence of word types in text corpora (Salton and McGill, 1983; Crouch, 1990; Chen *et al.*, 1995; Schütze and Pedersen, 1997). The co-occurrence method uses the collection of documents as a source for thesaurus construction according to a three-step procedure:

(1) Automatic identification of concepts within a given domain.

(2) Extraction of word types from the text.

(3) Cluster analysis to form possible thesaurus classes (Salton and McGill, 1983).

Using a collection of documents as a source for thesaurus construction entails that a representative body of text is available for application of statistical procedures to identify important terms as well as their significant relationships (Srinivasan, 1992). The assumption behind statistical methods is that contextually related co-occurring words (i.e. often appearing in the same sentence, paragraph, or document) are semantically related and hence should be classified together in the same thesaurus class (Srinivasan, 1992). The most commonly used method to automatic construction of thesauri is Salton's vector space model and term discrimination theory (Salton and McGill, 1983). Based on cluster analysis of terms in documents, the model has been extended from automatic indexing to automatic thesaurus construction (see section five for a description of matrix generation, similarity measures, and cluster analyses). The problem with the vector space model is the inherent dimensionality problem, which makes it computer intensive, and applicable only to smaller collections. In addition, problems exist as to the difficulty in identifying synonyms within the same documents, and the ever-present semantic problems caused by automatic indexing (Peat and Willett, 1991; Schütze and Pedersen, 1997; Lancaster, 1998). Another basic problem is to define the threshold value to determine the actual classifications (Crouch, 1990).

The purpose of automatically constructed thesauri, apart from making the construction and maintenance processes less resource demanding, is to improve retrieval performance by substituting the appropriate cluster of terms for one of its members (Salton and McGill, 1983). The classes formed by statistical procedures will tend to contain relatively more semantically different terms than those of a conventional thesaurus due to stemming procedures (Lancaster, 1998). According to Lancaster (1998, p. 263), the "purity" of the class is not always the main issue. It is important whether the class is potentially useful to IR. The heterogeneous nature of the clusters makes it more likely that recall rather than precision will be enhanced (Crouch, 1990; Chen *et al.*, 1995).

Statistically-based thesaurus construction may yield acceptable results when constructed from a large corpus of text with a specialized vocabulary, but the technique is questionable with heterogeneous text corpora (Salton and McGill, 1983). Moreover, the technique simply detects terms and possible relationships (most likely related terms). Detecting the specific semantic nature of these terms, and their relationships, is usually beyond their scope.

Grefenstette (1994) labels the different statistical co-occurrence methods as *first-order word associations* because they group together words that are recurrently found close to one another, in no particular order. In contrast, *second-order word associations* group words with similar contexts.

By means of syntactic analysis, one can determine which word is referring to which other word in a sentence. Term comparison based on these syntactic relations then leads to linguistically-based second-order associations (Hindle, 1990; Ruge, 1992; Grefenstette (1994); Jing and Croft, 1994). By contrast to statistically-based associations, linguistically-based associations are semantically compatible, therefore, one would suppose that the linguistically-based second-order associations are more semantically similar than the statistical ones (Ruge, 1992). Several different linguistically-based corpus methods exist. The difference lies in the syntactic relation extracted from the corpus. The most widely used syntactical relation within automatic thesaurus

construction is the head-modifier analysis of noun phrases. The modifier is the word that specifies the head. Each noun (head) has a set of verbs, adjectives and nouns that it co-occurs with (modifiers), a mutual information value is calculated for each using typically a log function (Mandala *et al.*, 1999). Finally, a similarity coefficient (typically cosine) between the words is calculated using the mutual information in order to classify the words. Syntactical context similarity analyses are used to locate synonyms (Grefenstette, 1994), as an indicator of hyponym/hypernym relationships (Hearst, 1998), or to determine concept hierarchies (Woods, 1997; Sanderson and Croft, 1999).

Second-order associations can be seen as an extension of first-order associations because they exploit classes of syntactic contexts instead of co-occurrence contexts. However, there is a limit to the quality of relations that can be achieved by pure automatic domain independent processing techniques (Kowalski and Maybury, 2000). Thus, automatic thesaurus construction methods cannot function alone, if an elaborate structure and semantic term validity is desired (Lancaster, 1998).

Bibliometrics applied to indexing, domain visualization, and thesaurus maintenance

Citations are commonly conceptualized as representations of the relationships between documents. A conceptualization of this kind can be characterized as being *artefact oriented*, because it is the citing/cited document pair (rather than the citing author) that is the dominant component of the definition. Further, studies that conceptualize citations as artefacts are typically more concerned with properties of documents rather than properties of people (Borgman and Furner, 2002). One such property is relatedness of content, the content of the earlier cited document is relevantly related to that of the later citing document (e.g. White and Wang, 1997). Grouping documents according to their relatedness of content by use of citation analysis links bibliometrics to other automatic algorithmic methods to construction of knowledge organization systems; as the automatic algorithmic approach and bibliometrics share the same epistemological foundation, *empiricism* to knowledge organization (e.g. Hjørland, 1997, 1998). That is, they employ analogous methods, based on statistical frequencies, but as pointed out previously, they apply different units of analysis. The traditional automatic algorithmic approach depends on the frequency distribution of index terms as units, where as bibliometrics depend on frequency distribution of citations and bibliographic references in documents. Thus, data computed within bibliometrics are based on sociological patterns of explicit recognition between individual documents, rather than statistical patterns of terms in documents. This also links bibliometrics to *pragmatism* (Hjørland, 1997, 2002b).

Citation indexing, mapping, and visualization can be looked on as knowledge organization. The use of bibliometric methods as a tool for construction of knowledge organization systems is not a new idea. Citation indexing can be traced back to 1873 when Shepard's citation index was established within the legal domain (Adair, 1955). It was not until the early 1950 s that the concept was conceived as a way to monitor, organize, and retrieve published scientific literature (Garfield, 1979). The Science Citation Index® (SCI®) launched by the Institute for Scientific Information®(ISI®) in the early 1960 s, is one of the first applications of computers in the production of large-scale, machine-generated indexes. Later Social Science Citation Index® (SSCI®) and Arts and Humanities Citation Index® (AandHCI®) have followed as well as Web of Science® all commercial indexes owned by ISI®.

Automatic indexing by use of citation linkages

The idea behind scientific citation indexes is linked to IR, i.e. related documents can be grouped on the basis of direct citation, co-citation, or bibliographic coupling as well as through the more conventional methods of subject indexing (e.g. Garfield, 1955). Kwok (1985a; 1985b) points to the fact that reference and citation linkages can be used in IR to form an "augmented collection" of retrieved items. This means that a set of items retrieved the conventional way, by use of text words or controlled terms, can be augmented by those items linked to them through bibliographic references and citations. Comparative studies of conventional subject indexing and citation linkage find that the two methods are essentially complementary (e.g. Kessler, 1965; Pao, 1988; Pao and Worthen, 1989; Shaw, 1990).

KeyWords Plus® is automatic indexing in which candidate terms, extracted from titles in the cited references of source documents, augment or replace author-designated keywords to documents in ISI's® citation and Current Contents® databases (Garfield, 1990). Recent applications of citation indexing include ResearchIndex (also known as CiteSeer, http://citeseer.nj.nec.com/), that is, a Web-based citation database. ResearchIndex uses citation indexing to allow users to search the Web for scientific documents and to retrieve citation contexts (Lawrence *et al.*, 1999).

Visualization studies

In relation to bibliometrics, visualization is often associated with concepts such as field, discipline, specialty area, knowledge structure, and knowledge domains (e.g. White and McCain, 1997; Chen, 2003; Börner *et al.*, 2003). In effect, these concepts vary only with respect to the focus of the research since the methods applied are basically the same, and the result is a visualization of structures and relationships within a subject literature.

Overall, visualization aims to capture perceived topical or intellectual structures (scientific communication) of a particular knowledge domain(s), reflected through the scientific literature, for example through its citation paths (e.g. Garfield *et al.*, 1964; White and McCain, 1997). Visualization means visual appearance of data objects and their relationships. This can be done manually more or less primitive but today visualization studies involve computation of large volumes of data and two or three-dimensional interactive spatial map representations (Börner *et al.*, 2003). The progress and growth of visualization studies is strongly connected to the advancements within computer technology and research in IR (White and McCain, 1997). There is an acceptance that visualization helps an increasingly diverse community to gain overviews of (general) patterns and trends, and to discover hidden semantic structures (Börner *et al.*, 2003). Domain visualization and domain analysis are associated in that the former may provide enabling techniques needed in domain analysis (Hjørland, 1997, 2002a; White and McCain, 1998; Boyack *et al.*, 2002; Chen *et al.*, 2001; Chen *et al.*, 2002). Essentially, visualizations in this context are visual reflections of specific knowledge organization attempts, and can be viewed as knowledge organization systems.

By use of bibliometric mapping knowledge structures in a given domain can be visualised. For a successive number of years ISI® identified "research fronts" for specialty areas in SCI® (Garfield, 1994). A "research front" is a group of current papers that cite a cluster of older "core" papers in a speciality area (Price, 1965).

Co-citation analysis is used in Atlas of Science® to make multiple levels of clusters, subsequently used for generation of "nested maps" that provide hierarchical or regionalized structures of "research fronts" in fields such as biochemistry or biotechnology (Garfield, 1994). Individual "research fronts" generally correspond to subspecialty areas, and multiple "research fronts" may represent the slight variations in the same subspecialty area (Garfield, 1994). It is possible to group related "research fronts" together to form higher-level aggregations that correspond to specialty areas. More recently, ISI® has developed the SCI-Map® software, which enables users to navigate a citation network (Small, 1999). Garfield (1994) has introduced the concept of longitudinal mapping where a series of chronological sequential maps can be used to detect the advances of scientific knowledge over time.

Researchers in the Netherlands have developed bibliometric mapping further, e.g. through self-organized structuring of scientific fields etc. (e.g. Braam *et al.*, 1991a, 1991b; Noyons and Van Raan, 1998a, 1998b, 1998c; Noyons *et al.*, 1999). Common for these attempts are that they identify sub-domains, within a research field for different time periods, using mapping and clustering techniques, as well as co-word analysis done on different content indicators such as noun phrases from titles and abstracts (e.g. Noyons and Van Raan, 1998a). This enables monitoring of scientific fields over time.

In common to these bibliometric attempts of knowledge organization are that they all rely fully on automatic algorithmic approaches. None of the citation indexing methods deal with either the inherent problems of indexing languages, such as terminological and semantic issues, or the possible relationship types between indexing terms in the language (Lancaster, 1998). Likewise, visualization studies are often defective when it comes to the interpretation of the results, or rather the non-existing interpretation of the result. This often generates debate focusing on what/who is missing on the maps, or the spatial positioning of the included objects.

In order to acknowledge these and related problems, we suggest a combination of bibliometric and intellectual manual construction methods. This may counter problems related to automatic algorithmic construction methods and vice versa, ensure a better understanding of the premises for construction, and creation of a "richer" context enabling a potential better interpretation of results obtained.

Semi-automatic thesaurus maintenance study

The research by Rees-Potter (1987, 1989) on semi-automatic thesaurus maintenance is an excellent example of a combination of manual intellectual and automatic algorithmic approaches. The objective of her study is to identify conceptual changes in two domains (sociology and economics) over time, as well as to investigate bibliometric methods' ability to identify candidate thesaurus terms (Rees-Potter, 1989). The employed bibliometric methods are citation analysis, co-citation analysis and citation context analysis. The results indicate that highly cited and co-cited documents act as concept symbols verifying former research by Small (1978) and Cozzens (1982). Thus, citing documents' citation context can be investigated for candidate thesaurus terms. However, in the case of Rees-Potter the investigated citation contexts primarily come from monographs due to domain specific conditions. None of the employed methods by Rees-Potter accomplish to identify conceptual changes over time. In addition, the applied citation context analysis is time consuming since it is done

manually. The initial intention of Rees-Potter was to implement her findings in a full-text system for semi-automatic thesaurus construction and maintenance, consequently, this has not been achieved. Overall, the research by Rees-Potter (1987, 1989) indicates the value of bibliometric methods for selection of candidate thesaurus terms beyond the traditional term co-occurrence methods. Some of the conditions and problems verified in her study are addressed in the present research project, and outlined in the following section that discusses appropriate bibliometric methods.

Methodical considerations concerning the suggested bibliometric based semi-automatic approach

This section concerns the methodical considerations of the proposed bibliometric based semi-automatic approach. Overall, the semi-automatic approach consists of four overlapping steps:

(1) Generation of a (reliable) text corpus (sampling).

(2) Structuring of the text corpus by use of different bibliometric methods (citation and co-citation analysis), cluster analysis and network analysis.

(3) Identification and extraction of candidate thesaurus terms from citation contexts by use of citation context analysis and syntactical parsing.

(4) Construction and visualization of conceptual networks coming from the extracted candidate thesaurus terms by use of bibliometric methods (co-word analysis), multivariate statistical analysis (multidimensional scaling), and network analysis.

The related work, presented above, serves as an inspiration for the choice of components incorporated into the semi-automatic approach for thesaurus construction and maintenance. Eventually, the proposed approach is applied in a case study to investigate the applicability of bibliometric methods for thesaurus construction and maintenance. Below, we present and discuss the essential aspects of the approach with respect to generation of a text corpus (sampling method), bibliometric methods (citation and co-citation analysis, citation context analysis, and co-word analysis), multivariate statistical techniques (cluster analysis and multidimensional scaling), and network analysis (Pathfinder Network Scaling). As mentioned above the approach is not definitive, other methods and techniques may be employed as well.

Generation of a text corpus – sampling

The first step concerns the generation of a text corpus. Traditional automatic thesaurus construction methods use the whole database as its text corpus not relying on document structure, such as paragraphs or references. As reported in Schneider and Borlund (2002) the text corpus is generated by use of the *data set isolation method* (Ingwersen and Christensen, 1997). The data set isolation method creates overlapping document sets at various points in time, i.e. the overall text corpus consist of four subsets (document sets). The subject area for the present research project is *periodontics*, a sub-domain to dentistry. The overlapping sets of documents are created through the merger of document representations from MEDLINE as well as SCI®, for four time periods. Data extraction from MEDLINE is chosen in order to utilize Medical Subject Headings (MeSH), titles and abstracts from the documents, and extraction from SCI® to "access" the same documents' bibliographic references.

We work with different time periods (in this case four time periods: 1989, 1993, 1997, and 2001) because we are interested in verifying whether bibliometric methods can be used for identification of change in terminology over time within a given domain, that is, handling the dynamic nature of language. The use of a collection of documents as a source for thesaurus construction entails that a representative body of text is available for application of statistical procedures to identify important terms as well as their significant relationships (Srinivasan, 1992).

The preliminary experiment yields positive results, that is, the overlapping document sets represent a solid sample of documents dealing with a variety of aspects of periodontics in the given time periods – indicating a representative body of text available for bibliometric analyses (Schneider and Borlund, 2002).

Bibliometric methods

The purpose behind structuring the text corpus through bibliometric methods is to identify topically related "core documents" and "core document groupings". Structuring of the text corpus starts with the choice of units of analysis. The most common used units in bibliometric analyses are journals, documents, authors and descriptive terms (White and McCain, 1997). Each unit presents a different facet of a domain and enables different analyses. Documents (and their concepts) are of interest to this study since ultimately they are the objects from where candidate thesaurus terms are extracted and relations uncovered. Documents are preferred as units for analysis when visualizing topical structures of knowledge domains (Börner *et al.*, 2003); where as author units (oeuvres) are typically used to infer intellectual structures of a field (White and McCain, 1998).

Next step, based on the units of analysis, is to choose what methods should be employed for the structuring the text corpus. In brief, bibliometric methods can be divided into two groups:

(1) Direct linkages such as direct citation.

(2) Indirect linkages such as co-citation or bibliographic coupling.

Bibliometric methods quantify similarities or dissimilarities between units hereby revealing structures and relationships. White and McCain (1997) note that bibliometric methods cover technical terms such as inter-citation, inter-document, co-assignment, co-classification, co-citation, and co-word analysis. "Inter" refers to relationships between documents (or units), and "co" refers to joint occurrences within a single document (or unit). Irrespective of the methods, applied counting may necessitate thresholds (Wilson, 1999). Thresholds are necessary for the creation of a data set of a manageable size, both in respect to computation and visualization (Börner *et al.*, 2003). This often causes debate because thresholds are artificial but critical to the mapping process (Van Leeuwen *et al.*, 1999). The rationale for using indirect linkages is that they reinforce regions of dense direct citation and thereby facilitate the down breaking of the citation network into meaningful chunks (Small, 1999).

As outline above, we investigate the following bibliometric methods' ability to identify structure, concepts, and relationships: citation analysis; co-citation analysis, citation context analysis, and co-word analysis. The rationale and methodical considerations of the listed methods are stated below.

Citation analysis (direct linkage) is applied to identify highly cited documents – "citation classics" – within the four documents sets (different time periods). Subsequently, the highly cited documents are subjected to co-citation analysis (indirect linkage), in order to illustrate the "research fronts" in the individual document sets, as well as to structure their "intellectual base" (Persson, 1994). In this context, empirical support for the application of citation analysis is provided by Small (1978) who established that cited documents symbolize concepts to those who cite them. The highly cited book by Kuhn (1962) may be perceived as a concept symbol referring to the concept of paradigm theory (Rees-Potter, 1989). Small (1978) documents that scientists tend to assign earlier works consensual meaning by "pilling up" identical or similar phrases in which their citation markers are embedded. Rees-Potter (1987, 1989) validates that concept symbols can be treated as candidate thesaurus terms. In addition, O'Conner (1983) applies automatic identification and tagging techniques to isolate citation contexts in order to extract single terms for automatic indexing.

The most commonly used bibliometric method for mapping and visualization studies is the co-citation analysis (Small, 1973; Small and Griffith, 1974; McCain, 1990; White and McCain, 1998). Co-citation analysis is generally accepted as a good indicator for illustrating "research fronts" (White and McCain, 1997). Co-citation analysis is based on two assumptions:

(1) When two documents are cited together by a third document, then a cognitive relationship exists between them.

(2) The strength of this relationship is proportional to the frequency of the co-citation linkage, i.e. the number of documents that co-cite the two documents.

Clusters of related documents can be constructed for a specified threshold of co-citation, see the section below on clustering. The relationships between clusters can be spatially displayed by use of, e.g. multidimensional scaling, see below. The clusters represent topics, specialities, or fields, while links between them reveal possible relationships (McCain, 1990).

Consequently, document co-citation analysis and mapping for the four time periods shows clusters ("research fronts") of co-cited documents and possible linkages between them. The co-citation maps are used to identify examples of core documents (potential concept symbols), "research fronts", possible topical structures, and relationships between clusters.

Semi-automatic citation context analysis is performed on a sample of the *citing*documents (i.e. documents that point to a cited document in a "research fronts") to establish if there is consensus on the cited document being a "concept symbol", i.e. terminological agreement. Noun phrases are automatically extracted from the context that surrounds the potential "concept symbol" (including the "concept symbol" itself) in citing MEDLINE documents. Noun phrases are chosen as principal terms due to their appropriateness in thesaurus construction (Soergel, 1974; Rees-Potter, 1989; Aitchison *et al.*, 2000). The result is a "*concept symbol word profile"*, that is, a concept symbol and a number of frequently occurring words and phrases attached to it. Thus, instead of using unstructured document text for thesaurus construction, we apply document structure in the form of citation contexts to identify candidate thesaurus terms.

"Research front" clusters are investigated for potential conceptual changes during the period under examination. This is done by a comparison between the concept

symbol word profiles attached to the individual "research front" clusters at different time periods (Braam *et al.*, 1991a, 1991b). Major changes in word profiles may indicate a shift in terminology used and in time a conceptual change. Comparison of concept symbol word profiles for the four time periods is intended to be able to investigate conceptual changes as expressed in the text of the documents. Finally, "concept symbol word profiles" for cited documents in a "research front" cluster is transformed into conceptual networks in order to investigate various conceptual relations between its terms (i.e. equivalence, hierarchical and associative relationships). Transformation is done by use of co-word analysis, multidimensional scaling, and network analysis.

Traditionally, the primary purpose of co-word analysis is for researches to analyse the dynamics of science and technology. Co-word analysis is inspired by the actor-network theory which fundamental premises is based on scientists' use of scientific publications as a vehicle for research ideas, hence creating a semiotic network of concepts (Callon *et al.*, 1983; Leydesdorff, 1997). A co-word analysis measures the strength of relationship between two documents by the co-occurrence of the same "words" (phrases, descriptors, classification codes etc.) in a chosen field. In co-word analysis, documents typically denote title, abstract, and/or descriptor fields (Callon *et al.*, 1983). However, in our case, documents are represented by the sample of citing documents' citation contexts. When the relationship strength between all pairs of documents has been calculated, multivariate statistical techniques can be applied to determine and map the required structure(s). Units of analysis connected to co-word analysis, i.e. words, phrases and descriptors may illustrate cognitive structures of a field when displayed in so-called "semantic maps" (Callon *et al.*, 1983; Braam *et al.*, 1991a, 1991b; Leydesdorff, 1997; Noyons and Van Raan, 1998a, 1998b, 1998c). We modify this conception in that specifically identified candidate thesaurus terms are used for co-word analysis in order to generate conceptual networks.

Two important issues in relation to co-citation and co-word analysis are the *similarity measures* used between units and *ordination*, i.e. the process of dimensionality reduction. The following sections will depict some of the characteristics that pertain to these issues.

Similarity measures

The various bibliometric methods measure relationships between units (e.g. documents, authors, words etc.). These measures are the basis for compilation of raw data matrices (McCain, 1990). The construction of a matrix with the same variable on both margins, and the cells containing the count (full or fractional) of either documents, or of links common to row-column pairs, are called co-matrices (e.g. co-citation matrices, co-word matrices). It is desirable to convert the raw data matrix into a matrix of proximity values, which indicate the relative similarity or dissimilarity of the units investigated. The creation of a proximity matrix has at least two major advantages:

(1) The similarity coefficient functions as a measure of pairwise similarity, not just a raw count

(2) It enables normalization procedures (e.g. Small and Greenlee, 1980).

To transform the raw data co-unit matrix into a proximity matrix requires a sequence of decisions about appropriate similarity measures. The first decision is the selection of

a threshold necessary to ensure that an adequate similarity structure will be extracted from the raw data matrix (Small and Greenlee, 1980; Braam *et al.*, 1988).

The second decision concerns the choice of indices to measure the degree of similarity from the raw counts. Different indices can differ in properties, such as normalization. The different properties of the indices affect the result of the clustering to varying degrees, as well as any later mapping (Börner *et al.*, 2003). The Extended Jaccard and the cosine indices are most commonly used in document co-citation (Small and Greenlee, 1980). Pearson's product moment correlation coefficient is mainly used in author co-citation analysis, and the Inclusion and extended Jaccard indices are typically used in co-word analyses (Leydesdorff, 1997; White and McCain, 1998). Braam *et al.* (1988) recommend to employ the simple cosine and Jaccard indices in parallel analysis. The choice of indices relates to the intended application of the bibliometric methods, and their resulting co-matrices (Leydesdorff, 1987). We use the Extended Jaccard index as similarity measure for practical reasons, since this is the preferred index for document co-citation clustering used by ISI® (Small and Greenlee, 1980). Further, we apply both the Inclusion and Extended Jaccard indices for the co-word analyses, since they essentially depict different relational aspects between the terms in the conceptual network.

The third decision regards the choice of how to cluster the transformed data (McCain, 1990) – this is outlined below.

Ordination

A major problem in relation to visualization of multivariate data, is the problem that they can be displayed only on a two or three-dimensional surface and with limited resolution. An alternative to this approach is the use of network analysis explained below. In cases with data containing more than three dimensions the problem is attempt solved by dimensionality reduction algorithms in order to map n-dimensional data into a two or three-dimensional space. The purpose of dimensionality reduction algorithms is to place objects that are similar to one another in n-dimensions close to one another and to place dissimilar objects far apart. This process is called ordination. The most common used multivariate statistical techniques for ordination are cluster analysis, multidimensional scaling and factor analysis (principal components analysis). The rationale and methodical considerations of the ordination techniques chosen for the semi-automatic approach are stated below, one by one.

Cluster analysis

A cluster analysis is a statistical technique used to generate a category structure (group) that fits a set of observations. The groups formed should have a high degree of association between members of the same group, and low degree of association between members of different groups (Everitt, 1998). The generated clusters (groups) are not known prior to processing but are defined by the objects assigned to them. Because there is no need for the clusters to be identified prior to processing, cluster analysis is useful to provide structure in large multivariate data sets (Everitt, 1998). Kowalski and Maybury (2000, p. 140) describe the process of cluster generation according to four steps:

(1) Define the domain for the clustering, i.e. identification of objects (documents) to be used in the clustering process and reduce potential for erroneous data.

(2) Determine the attributes of the objects to be clustered, e.g. documents' references and citations may be used to determine subject relatedness.

(3) Determine the strength of the relationships between the attributes whose co-occurrence in documents suggest those documents should be in the same group (similarity measures and thresholds).

(4) Applying a clustering algorithm to determine the cluster(s) to which each document will be assigned.

Many different clustering methods are available with different theoretical or empirical bases and they therefore produce different cluster structures (Börner *et al.*, 2003). For a given clustering method, there may be a choice of clustering algorithm to implement the method. The choice of clustering method determines the outcome; the choice of algorithm determines the efficiency with which it is achieved (Everitt, 1998). Clustering methods are usually categorized according to the type of cluster structure they produce. The simple non-hierarchical methods divide the data set of N objects (documents) into Mclusters, where no overlap is allowed, simple non-hierarchical methods are known as partitioning methods. Each object has a membership in the cluster with which it is most similar, and the cluster may be represented by a "centroid" or "cluster representative" that is indicative of the characteristics of the objects it contains (Kowalski and Maybury, 2000). This technique has been applied in several automatic thesaurus construction attempts (Salton and McGill, 1983).

More complex hierarchical methods produce a nested data set in which pairs of objects are successively linked until every object in the data set is connected. The hierarchical methods can be either agglomerative with $N-1$ pair-wise joins beginning from an un-clustered data set, or divisive beginning with all objects in a single cluster and progressing through $N-1$ divisions of some cluster into a smaller cluster (Kowalski and Maybury, 2000). The divisive methods are less commonly used in IR and bibliometrics, here the preferred clustering approach is the hierarchical agglomerative (Rasmussen, 1992).

The cluster structure, as the result of a hierarchical agglomerative clustering method, is often displayed as a dendogram or circle plot. The most common hierarchical agglomerative clustering algorithms are single link, complete link, and Ward's algorithm (Han and Kamber, 2000). The algorithms differ primarily in how similarity is defined. Each algorithm will produce different sets of clusters based on the same proximity matrix (Han and Kamber, 2000). The best algorithm is the one that consistently performs well on the co-occurrence data in terms of providing interpretable results (Han and Kamber, 2000). Traditional co-citation analysis relies on simple single-linkage clustering, because of its lower computational complexity given the typically large number of documents in a collection. For example, the approach to document co-citation clustering at ISI® has been the simple linkage algorithm (Small, 1973; Small and Griffith, 1974; Small and Greenlee, 1980). However, researchers concede a weakness in the single-linkage clustering approach (Small, 1993). The concern is with the possible "chaining" effect, in which unrelated documents are clustered together through a chain of intermediate documents. That is, a document need only be sufficiently co-cited with a single member of a cluster to be included in that cluster. Conversely, complete-linkage clustering is the strongest of the traditional graph-theoretic clustering methods (Han and Kamber, 2000).

In complete-link clustering a document must be sufficiently co-cited with all other cluster members to be included in the cluster. Such a strong clustering criterion insures that documents within a cluster are directly rather than indirectly related to one another (Han and Kamber, 2000). The resulting smaller, more cohesive clusters should better support micro-scale studies of document collections. However, the strong clustering criterion that pertains to the complete-linkage algorithms can result in a number of singletons that do not cluster at some threshold since their co-citations are distributed between other variables in the matrix, in such a way that no significant similarity is obtained.

We apply a heuristic agglomerative clustering approach. Initially, highly cited documents in the different document sets are chosen as objects for clustering. The co-citation normalization procedure, established in the proximity matrix (Extended Jaccard), reveals if some of the documents are not suitable for complete-link clustering. That is, they are likely to become singletons. Small and Greenlee (1980) have pointed out that such documents are typical of a methodical nature.

The key to successful clustering lies in the selection of a good similarity index and selection of a good clustering algorithm for placing objects in the same group, this is typically achieved through iterative trial and error attempts (Han and Kamber, 2000).

Multidimensional scaling (MDS)

MDS requires as input the same proximity matrix of similarities or dissimilarities among objects as cluster analysis. MDS is a set of techniques used to create visual displays (maps) from proximity matrices, so that the underlying structure within a set of objects can be studied (Kruskal, 1977). The result is a least-square representation of the objects by use of scatter plots. The plots are represented on a two or three-dimensional map according to their proximity in the original matrix transformed by the MDS program to a table of spatial coordinates (White and McCain, 1997). A major purpose of MDS is to capture as much of the original data as possible in only two or three dimensions, i.e. to reduce space. This simplification is valuable, but necessarily distorts the original data somewhat and cannot account for all the variances in the proximity matrix. MDS programs summarize the distortion with a statistical technique called "stress". The "stress" value is a criterion for determining the optimal match between the distances in the original matrix and the estimated distances in the chosen low-dimensional solution. The "stress" value is the indicator of the overall goodness-of-fit of that plot configuration (Börner *et al.*, 2003). MDS is one of the most popular mapping techniques in bibliometrics applied to, e.g. author co-citation analysis (White and McCain, 1998), science mapping (Small, 1999), and performance assessment (Noyons *et al.*, 1999).

We use the MDS-ALSCAL algorithm to produce two-dimensional maps for all generated matrices. The purpose of using MDS is to provide an information rich display of the various linkages between objects (co-citation and co-word), and to identify the salient dimensions underlying their placement. MDS maps done for all four time periods are investigated for visual changes in the maps, possibly indicating conceptual and/or structural changes in the domain requiring thesaurus maintenance. The main purpose of using MDS maps is to investigate the maps' intuitive and visual ability to reveal candidate thesaurus terms and conceptual relationships.

However, the use of MDS is not without problems, thus, we use network analysis to complement for these inexpediencies

Network analysis

Mathematically, networks can be represented as graphs and matrices (Scott, 2000). A graph consists of vertices and edges, in networks they are named nodes and links (Chen, 2003). Many important phenomena can be formulated as a graph problem including citation networks. In graph theory, the focus is on the connectivity of a graph: the topology rather than the geometry (Chen, 2003). The latter is the focus in MDS and factor analysis.

Network analysis is the mapping and measuring of relationships and flows between nodes in a network (Scott, 2000). Social networks, for example, are graphs in which nodes represent people and edges represent interrelationships between people. Network analysis allows relational structures inherent in data matrices to be investigated. A co-citation network is typically expressed as a relational data matrix. Network analysis is therefore an obvious investigational tool that can create simplicity and clarity of the hidden structures embedded in such a data matrix. Several metrics can be applied to network analysis. The most widely used metric is the centrality measure of node (Scott, 2000). The most popular centrality measures are "degrees", "betweenness", and "closeness". These measures help determine the importance, or prominence, of a node in the network (Scott, 2000). Other network metrics include structural equivalence, which determines which nodes play similar roles in the network, cluster analysis that identify cliques and other densely connected clusters, structural holes, which identify areas of no connection between nodes etc. (Scott, 2000).

Pathfinder Network Scaling originally developed by cognitive psychologists for modelling networks of concepts (Schvaneveldt, 1990), have gained much attention within citation analysis in the last couple of years (Chen *et al.*, 2001; Chen *et al.*, 2002, Chen, 2003; White, 2003a). Pathfinder Network Scaling relies on a triangle inequality condition to select the most salient relations from proximity data (Chen, 2003). Pathfinder networks (PFNETs) have the same set of nodes as the original graph; however, the number of links in a Pathfinder network can be greatly reduced. Pathfinder Network Scaling selects "important" links into the final network representation (Chen, 2003). PFNETs are scale-free networks and the spatial layout is based on the spring-embedder algorithm where link distance is uniformly rendered, and void space has no semantics in its own right (Chen, 2003). Connectivity and paths are the predominate objects for interpretation. PFNETs display links between objects explicitly, and structural patterns are therefore relatively easy for our perception to detect.

Dimensionality reduction algorithms (e.g. MDS and factor analysis) can reduce implicit links – dimensions. Link reduction algorithms (e.g. minimum spanning trees and Pathfinder Network Scaling) can reduce explicit links – connections. At the same time, link reduction algorithms may indicate groupings due to connectivity between the nodes in the network.

Indeed, the use of PFNETs as an alternative to MDS in for instance author co-citation analysis seems to be very fruitful (White, 2003a). PFNETs avoid to a great extent some of the most fundamental problems with MDS maps: scalability; and MDS maps do not necessarily group explicit information together so that patterns must be

judged carefully to identify underlying structure. In addition, the recent debate on the validity of some similarity indices in relation to MDS (Ahlgren *et al.*, 2003; White, 2003a, 2003b) can to a large extent be avoided by using PFNETs based on raw co-citation counts, as shown by Chen (2003) and White (2003a).

We intend to use Pathfinder Network Scaling and network analysis as a supplement to the above mentioned dimensionality reduction algorithms. We wish to compare the two approaches to seek out their advantages and disadvantages in relation to analyzing both citation and conceptual networks. Specifically, we wish to identify the most salient and "important" connections in the network, they may indicate relationships between concepts and concept groups. Our main assumption is that network analysis and Pathfinder Network Scaling are simpler to interpret and much more clear in revealing structures for larger data sets than for instance traditional MDS.

Summary statements

Bibliometrics is epistemologically founded in empiricism, as well as pragmatism, and is thereby linked to the automatic algorithmic approach to knowledge organization. However, the application of bibliometric methods is not a completely automatic thesaurus construction process, as intellectual interpretation of identified candidate terms and concept relationships is required. Hence, we refer to the described approach as a semi-automatic approach to thesaurus construction and maintenance.

The proposed approach is substantially different from the traditional automatic algorithmic term co-occurrence methods to thesaurus construction; because term class construction in the bibliometric approach is based on citations and references. Citations and references are independent of language and changing terminology, and consequently this can be exploited advantageously for thesaurus construction and maintenance purposes. Thus, bibliometric methods are recommendable for uncovering different knowledge patterns in texts, through the use of citations and references given in the scientific literature, as a supplement to the traditional approaches.

Our approach and introduction of bibliometric methods to thesaurus construction and maintenance is new. The application of bibliometrics is rarely seen, with the exception of Rees-Potter (1987, 1989), who supports our approach. The proposed range of bibliometric methods is commonly known. The proposed range of methods and techniques to be verified are not definitive, but represent our choices of methods based on the presented methodical considerations. Further, the introduction of bibliometrics is not restricted to thesaurus construction and maintenance, but is also applicable to less elaborated and sophisticated knowledge organization systems, such as semantic networks and ontologies.

Two main approaches exist to knowledge organization, that is, the manual intellectual approach and automatic algorithmic approach. The manual intellectual approach is considered essential to construction of knowledge organization systems due to the dynamic and complex nature of language. But the manual intellectual approach is also acknowledged for being a resource demanding and costly process. In contrast, the automatic algorithmic approach is less resource demanding and is furthermore suitable for managing large data sets, what makes it an attractive approach to knowledge organization. Consequently, the two main approaches supplement one another, and can advantageously be combined to counterbalance

the strengths and weaknesses of each of them. Motivated by this bibliometrics is proposed as a semi-automatic approach to supplement the manual intellectual approach to thesaurus construction and maintenance – and is presented as the basic *idea* of the paper.

Notes

1. The authors use the term "documents" to denote all kinds of information objects or texts in a semiotic sense, i.e. organized sets of symbols chosen to represent a message – though the focus in the present paper concerns text documents.
2. Items refer to terms as well as documents.
3. However, declining citation rates also occur for papers containing important ideas. Once an idea is sufficiently widely known, citing the original version is unnecessary. This phenomenon has been termed 'obliteration by incorporation' (Garfield, 1975).

References

Adair, W.C. (1955), "Citation indexes for science", *American Documentation*, Vol. 6 No. 1, pp. 31-2.

Ahlgren, P., Jarneving, B. and Rousseau, R. (2003), "Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 6, pp. 550-60.

Aitchison, J., Gilchrist, A. and Bawden, D. (2000), *Thesaurus Construction and Use: A Practical Manual*, 4th ed., Aslib, London.

Anderson, J.D. (1997), "Organization of Knowledge", in Feather, J. and Sturges, P. (Eds), *International Dictionary of Library and Information Science*, Routledge, London, pp. 336-53.

Anderson, J.D. and Perez-Carballo, J. (2001a), "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: research, and the nature of human indexing", *Information Processing and Management*, Vol. 37 No. 2, pp. 231-54.

Anderson, J.D. and Pérez-Carballo, J. (2001b), "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: machined indexing, and the allocation of human versus machine effort", *Information Processing & Management*, Vol. 37 No. 2, pp. 255-77.

Blair, D.C. (1990), *Language and Representation in Information Retrieval*, Elsevier, Amsterdam.

Borgman, C.L. (1990), *Scholarly Communication and Bibliometrics*, Sage, London.

Borgman, C.L. and Furner, J. (2002), "Scholary communication and bibliometrics", in Cronin, B. (Ed.), *Annual Review of Information Science and Technology*, Vol. 36, (forthcoming).

Börner, K., Chen, C. and Boyack, K.W. (2003), "Visualizing knowledge domains", in Cronin, B. (Ed.), *Annual Review of Information Science and Technology*, Vol. 37, (forthcoming).

Boyack, K.W., Wylie, B.N. and Davidson, G.S. (2002), "Domain visualization using VxInsight for science and technology management", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 9, pp. 764-74.

Braam, R.R., Moed, H.F. and Van Raan, A.F.J. (1988), "Mapping of science: critical elaboration and new approaches, a case study in agricultural biochemistry", in Egghe, L. and Rousseau, R. (Eds), *Informetrics 87/88: Select proceedings of the 1st International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval,* 25-28, August, 1987, Diepenbeek, Belgium. Elsevier, Amsterdam, pp. 15-18.

Braam, R.R., Moed, H. and Van Raan, A.F.J. (1991a), "Mapping of Science by combined co-citation and word analysis. I. Structural aspects", *Journal of the American Society for Information Science*, Vol. 42 No. 4, pp. 233-51.

Braam, R.R., Moed, H. and Van Raan, A.F.J. (1991b), "Mapping of science by combined co-citation and word analysis. II. Dynamical aspects", *Journal of the American Society for Information Science*, Vol. 42 No. 4, pp. 252-66.

Burton, R.E. and Kebler, R.W. (1960), "The 'half-life' of some scientific and technical literatures", *American Documentation*, Vol. 11 No. 1, pp. 18-22.

Callon, M., Courtial, J.P., Turner, W.A. and Bauin, S. (1983), "From translation to problematic networks: an introduction to co-word analysis", *Social Science Information*, Vol. 22 No. 2, pp. 191-235.

Chen, C. (2003), *Mapping Scientific Frontiers – The Quest for Knowledge Visualization*, Heidelberg: Springer-Verlag, Berlin.

Chen, C., Cribben, T., Macredie, R. and Morar, S. (2002), "Visualizing and tracking the growth of competing paradigms: two case studies", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 8, pp. 678-89.

Chen, C., Paul, R.J. and O'Keefe, B. (2001), "Fitting the jigsaw of citation: information visualization in domain analysis", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 4, pp. 315-30.

Chen, H., Schatz, B., Yim, T. and Fye, D. (1995), "Automatic thesaurus generation for an electronic community system", *Journal of the American Society for Information Science*, Vol. 46 No. 3, pp. 175-93.

Cole, S. (1992), *Making Science. Between Nature and Society*, Harvard University Press, Cambridge, MA.

Cozzens, S.E. (1982), "Split citation identity: a case study from economics", *Journal of the American Society for Information Science*, Vol. 33 No. 4, pp. 233-6.

Cozzens, S.E. (1989), "What do citations count? The rhetoric first model", *Scientometrics*, Vol. 15 Nos 5/6, pp. 437-47.

Cronin, B. (1984), *The Citation Process: The Role and Significance of Citations in Scientific Communication*, Taylor Graham, London.

Crouch, C.J. (1990), "An approach to automatic construction of global thesauri", *Information Processing & Management*, Vol. 26 No. 5, pp. 629-40.

Egghe, L. and Rousseau, R. (1990), *Introduction to Informetrics*, Elsevier, Amsterdam.

Everitt, B.S. (1998), *Cluster Analysis*, 3rd ed., Edward Arnold, London.

Garfield, E. (1955), "Citation indexes for science: a new dimension in documentation through association of ideas", *Science*, Vol. 122 No. 3159, pp. 108-11.

Garfield, E. (1965), "Can citation indexing be automated?", in Stevens, M.R., Giuliano, V.E. and Heilprin, L.B. (Eds), *Statistical Association Methods for Mechanized Documentation*, NBS, Washington, DC, pp. 189-92.

Garfield, E. (1975), "The 'obliteration phenomenon' in science and the advantage of being obliterated!", *Current Contents*, Vol. 51/52, pp. 5-7.

Garfield, E. (1979), *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*, John Wiley & Sons Inc., New York, NY.

Garfield, E. (1990), "KeyWords plus", *Current Contents*, Vol. 32, pp. 3-7.

Garfield, E. (1994), "Research fronts", *Current Contents*, Vol. 41, pp. 3-6.

Garfield, E. (1997), "Validation of citation analysis", *Journal of the American Society for Information Science*, Vol. 48 No. 10, p. 962.

Garfield, E. (1998), "Random thoughts on citationology: its theory and practice", *Scientometrics*, Vol. 43 No. 1, pp. 69-76.

Garfield, E., Sher, I.H. and Torpie, R.J. (1964), *The Use of Citation Data in Writing the History of Science*, Institute for Scientific Information, Philadelphia.

Glänzel, W. and Schoepflin, U. (1994), "Little scientometrics – big scientometrics. . . and beyond", *Scientometrics*, Vol. 30 Nos 2/3, pp. 375-84.

Grefenstette, G. (1994), *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publisher, Boston, MA.

Han, J. and Kamber, M. (2000), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Mateo, CA.

Hearst, M. (1998), "Automated discovery of WordNet relations", in Fellbaum, C. (Ed.), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Hindle, D. (1990), "Noun classification from predicate argument structures", *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics held in Pittsburg, PA*.

Hjørland, B. (1997), *Information Seeking and Subject Representation. An Activity-Theoretical Approach to Information Science*, Greenwood Press, London.

Hjørland, B. (1998), "The classification of psychology: a case study in the classification of a knowledge field", *Knowledge Organization*, Vol. 24 No. 4, pp. 162-201.

Hjørland, B. (2002a), "Domain analysis in information science. Eleven approaches – traditional as well as innovative", *Journal of Documentation*, Vol. 58 No. 4, pp. 422-62.

Hjørland, B. (2002b), "The methodology of constructing classification schemes: a discussion of the state-of-art", in López-Huertas, M.J. (Ed.), *Challenges in Knowledge Representation and Organization for the 21th Century. Integration of Knowledge across Boundaries, Proceedings of the Seventh International ISKO Conference, 10-13 July 2002, Granada, Spain, Ergon Verlag, Würzburg*, pp. 450-456.

Ingwersen, P. and Christensen, F.H. (1997), "Data set isolation for bibliometric online analyses of research publications: fundamental methodological issues", *Journal of the American Society for Information Science*, Vol. 48 No. 3, pp. 205-17.

Jing, Y. and Croft, W.B. (1994), "An association thesaurus for information retrieval", *Proceedings of RIAO '94*, 11-13 October, New York, NY, pp. 146-160.

Kessler, M.M. (1963), "Bibliographic coupling between scientific papers", *American Documentation*, Vol. 14 No. 1, pp. 10-25.

Kessler, M.M. (1965), "Comparison of the results of bibliographic coupling and analytic subject indexing", *American Documentation*, Vol. 16 No. 3, pp. 223-33.

King, J. (1987), "A review of bibliometric and other science indicators and their role in research evaluation", *Journal of Information Science*, Vol. 13 No. 5, pp. 261-76.

Kostoff, R.N. (1998), "The use and misuse of citation analysis in research evaluation", *Scientometrics*, Vol. 43 No. 1, pp. 27-43.

Kowalski, G.J. and Maybury, M.T. (2000), *Information Storage and Retrieval Systems. Theory and Implementation*, Kluwer Academic Publishers, Norwell, MA.

Kruskal, J.B. (1977), "The relationship between multi-dimensional scaling and clustering", in Van Ryzin, J. (Ed.), *Classification and Clustering*, Academic Press, New York, NY, pp. 17-44.

Kuhn, T. (1962), *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, IL.

Kwok, K.L.A. (1985a), "A probabilistic theory of indexing and similarity measure based on cited and citing documents", *Journal of the American Society for Information Science*, Vol. 36 No. 5, pp. 342-51.

Kwok, K.L.A. (1985b), "A probabilistic theory of indexing using author-provided relevance information", *Proceedings of the American Society for Information Science*, Vol. 22, pp. 59-63.

Lancaster, F.W. (1998), *Indexing and Abstracting in Theory and Practice*, Library Association Publishing, London.

Latour, B. (1987), *Science in Action*, Open University, Milton Keynes.

Lawrence, S., Giles, C.L. and Bollacker, K. (1999), "Digital libraries and autonomous citation indexing", *IEEE Computer*, Vol. 32 No. 6, pp. 67-71.

Leydesdorff, L. (1987), "Various methods for the mapping of science", *Scientometrics*, Vol. 11 No. 5-6, pp. 281-320.

Leydesdorff, L. (1997), "Why words and co-words cannot map the development of the sciences", *Journal of the American Society for Information Science*, Vol. 48 No. 5, pp. 418-27.

Leydesdorff, L. (1998), "Theories of citation?", *Scientometrics*, Vol. 43 No. 1, pp. 5-25.

Leydesdorff, L. and Wouters, P. (1999), "Between texts and contexts: advances in theories of citation? (A Rejoinder)", *Scientometrics*, Vol. 44 No. 2, pp. 169-82.

Line, M. (1993), "Changes in the use of literature with time – obsolescence revisited", *Library Trends*, Vol. 41 No. 4, pp. 665-83.

Liu, M. (1993), "Progress in documentation: the complexities of citation practice – a review of citation studies", *Journal of Documentation*, Vol. 49 No. 4, pp. 370-408.

Luukkonen, T. (1997), "Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis", *Scientometrics*, Vol. 38 No. 1, pp. 27-37.

McCain, K. (1990), "Mapping authors in intellectual space: a technical overview", *Journal of the American Society for Information Science*, Vol. 41 No. 6, pp. 433-43.

MacRoberts, M.H. and MacRoberts, B.R. (1989a), "Problems of citation analysis: a critical review", *Journal of the American Society for Information Science and Technology*, Vol. 40 No. 5, pp. 342-9.

MacRoberts, M.H. and MacRoberts, B.R. (1989b), "Another test of the normative theory of citing", *Journal of the American Society for Information Science*, Vol. 16, pp. 151-72.

MacRoberts, M.H. and MacRoberts, B.R. (1996), "Problems of citation analysis", *Scientometrics*, Vol. 36 No. 3, pp. 435-44.

Mandala, R., Tokunaga, T. and Tanaka, H. (1999), "Combining evidence from different types of thesaurus for query expansion", *Proceedings of the 22nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA*.

Merton, R.K. (1973), *The Sociology of Science: Theoretical and Empirical Investigations*, University of Chicago Press, Chicago, IL.

Miller, U. (1997), "Thesaurus construction: problems and their roots", *Information Processing & Management*, Vol. 33 No. 4, pp. 481-93.

Narin, F., Olivastro, D. and Stevens, K.A. (1994), "Bibliometrics theory, practice and problems", *Evaluation Review*, Vol. 18 No. 1, pp. 65-75.

Noyons, E.C.M. and Van Raan, A.F.J. (1998a), *Mapping Scientometrics, Informetrics, and Bibliometrics*, available at: www.cwts.nl/ed/sib/home.html (accessed 9 January 2002).

Noyons, E.C.M. and Van Raan, A.F.J. (1998b), "Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research", *Journal of the American Society for Information Science*, Vol. 49 No. 1, pp. 68-81.

Noyons, E.C.M. and Van Raan, A.F.J. (1998c), "Advanced mapping of science and technology", *Scientometrics*, Vol. 41 No. 1-2, pp. 61-7.

Noyons, E.C.M., Moed, H.F. and Luwel, M. (1999), "Combining mapping and citation analysis for evaluative bibliometric purposes: a bibliometric study", *Journal of the American Society for Information Science*, Vol. 50 No. 2, pp. 115-31.

O'Connor, J. (1983), "Biomedical citing statements: computer recognition and use to aid full-text retrieval", *Information Processing and Management*, Vol. 19 No. 6, pp. 361-8.

Pao, M.L. (1988), "Term and citation searching: a preliminary report", *Proceedings of the American Society for Information Science*, Vol. 25, pp. 177-80.

Pao, M.L. and Worthen, D.B. (1989), "Retrieval effectiveness by semantic and citation searching", *Journal of the American Society for Information Science*, Vol. 40 No. 4, pp. 226-35.

Peat, H.J. and Willett, P. (1991), "The limitations of term co-occurrence data for query expansion in document retrieval systems", *Journal of the American Society for Information Science*, Vol. 42 No. 5, pp. 378-83.

Peritz, B.C. (1992), "On the objectives of citation analysis: problems of theory and method", *Journal of the American Society for Information Science*, Vol. 43 No. 4, pp. 448-51.

Price, D. De Solla (1965), "Networks of scientific papers", *Science*, Vol. 149, pp. 510-5.

Price, D. De Solla (1986), *Little Science, Big Science . . . And Beyond*, Columbia University Press, New York, NY.

Rasmussen, E. (1992), "Clustering algorithms", in Frakes, W.B. and Baeza-Yates, R. (Eds), *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, Upper Saddle River, NJ, pp. 419-42.

Rees-Potter, L.K. (1987), "A bibliometric analysis of terminological and conceptual change in sociology and economics with the application to the design of dynamic thesaural systems", 2 volumes, PhD dissertation, University of Western Ontario, Ontario.

Rees-Potter, L.K. (1989), "Dynamic thesaural systems: a bibliometric study of terminological and conceptual change in sociology and economics with the application to the design of dynamic thesaural systems", *Information Processing & Management*, Vol. 25 No. 6, pp. 677-91.

Persson, O. (1994), "The intellectual base and research front of JASIS 1986-1990", *Journal of the American Society for Information Science*, Vol. 45 No. 1, pp. 31-8.

Ruge, G. (1992), "Experiments on linguistically-based term associations", *Information Processing and Management*, Vol. 28 No. 3, pp. 317-32.

Salton, G. and McGill, M.J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY.

Sanderson, M. and Croft, W.B. (1999), "Deriving concept hierarchies from text", *Proceedings of the 22nd Annual ACM SIGIR Conference on Research in Information Retrieval,* 15-19 August, Berkeley, CA, 206-213.

Schvaneveldt, R.W. (Ed.), (1990), *Pathfinder Associative Networks: Studies in Knowledge Organization*, Ablex Publishing Corporation, Norwood, NJ.

Schneider, J.W. and Borlund, P. (2002), "Preliminary study of the potentiality of bibliometric methods for the construction of thesauri", in Bruce, H., Fidel, R., Ingwersen, P. and Vakkari, P. (Eds), *Proceedings of the Fourth International Conference on Conceptions of*

Library and Information Science, Seattle, WA, July 21-25, Libraries Unlimited, Greenwood Village, CO, 151-165.

Schütze, H. and Pedersen, J.O. (1997), "A cooccurrence-based thesaurus and two applications to information retrieval", *Information Processing & Management*, Vol. 33 No. 3, pp. 307-18.

Scott, J. (2000), *Social Network Analysis – A Handbook*, 2nd ed., Sage Publications, London.

Seglen, P.O. (1998), "Citation rates and journal impact factors are not suitable for evaluation of research", *Acta Orthopaedica Scandinavica*, Vol. 69 No. 3, pp. 224-9.

Shaw, W.M. Jr (1990), "Subject indexing and citation indexing", *Information Processing & Management*, Vol. 26 No. 6, pp. 693-718.

Small, H. (1973), "Co-citation in the scientific literature: a new measure of the relationship between two documents", *Journal of the American Society for Information Science*, Vol. 24 No. 4, pp. 265-9.

Small, H. (1977), "A co-citation model of a scientific speciality: a longitudinal study of collagen research", *Social Studies of Science*, Vol. 7, pp. 139-66.

Small, H. (1978), "Cited documents as concept symbols", *Social Studies of Science*, Vol. 8, pp. 327-40.

Small, H. (1982), "Citation context analysis", in Dervin, B. and Voigt, M.J. (Eds), *Progress in Communication Sciences*, Vol. 3, Ablex, Norwood, NJ, pp. 287-310.

Small, H. (1993), "Macro-level changes in the structure of co-citation clusters: 1983-1989", *Scientometrics*, Vol. 26 No. 1, pp. 5-20.

Small, H. (1998), "Citations and consilience in science", *Scientometrics*, Vol. 43 No. 1, pp. 143-8.

Small, H. (1999), "Visualizing science by citation mapping", *Journal of the American Society for Information Science*, Vol. 50 No. 9, pp. 799-813.

Small, H. and Greenlee, E. (1980), "Citation context analysis of a co-citation cluster: recombinant DNA", *Scientometrics*, Vol. 2 No. 4, pp. 277-301.

Small, H. and Griffith, B.C. (1974), "The structure of scientific literature. I: identifying and graphing specialities", *Science Studies*, Vol. 4 No. 17, pp. 17-40.

Smith, L.C. (1981), "Citation analysis", *Library Trends*, Vol. 30 No. 1, pp. 83-106.

Soergel, D. (1974), *Indexing Languages and Thesauri: Construction and Maintenance*, Melville, Los Angeles, CA.

Srinivasan, P. (1992), "Thesaurus construction", in Frakes, W.B. and Baeza-Yates, R. (Eds), *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, Upper Saddle River, NJ, pp. 161-218.

Száva-Kováts, E. (2002), "Unfounded attribution of the 'half-life' index-number of literature obsolescence to Burton and Kebler: a literature science study", *Journal of the American Society for Information Science*, Vol. 53 No. 13, pp. 1098-105.

Tague-Sutcliffe, J. (1992), "An introduction to informetrics", *Information Processing & Management*, Vol. 28 No. 1, pp. 1-3.

Van Leeuwen, T.N., Moed, H.F. and Reedijk, J. (1999), "Critical comments on Institute for Scientific Information impact factors: a sample of inorganic molecular chemistry journals", *Journal of Information Science*, Vol. 25 No. 6, pp. 489-98.

Van Raan, A.F.J. (1998), "In matters of quantitative studies of science the fault of theorists is offering too little and asking too much", *Scientometrics*, Vol. 43 No. 1, pp. 129-39.

Vinkler, P. (1998), "Comparative investigation of frequency and strength of motives toward referencing: the reference threshold model", *Scientometrics*, Vol. 43 No. 1, pp. 107-27.

White, H.D. (1990), "Author co-citation analysis: overview and defense", in Borgman, C.L. (Ed.), *Scholary Communication and Bibliometrics*, Sage Publications, Newbury Park, CA, pp. 84-106.

White, H. (2003a), "Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientists", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 5, pp. 423-34.

White, H. (2003b), "Author Cocitation analysis and Pearson's r", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 13, pp. 1250-9.

White, H.D. and Griffth, B.C. (1981), "Author co-citation: a literature measure of intellectual structure", *Journal of the American Society for Information Science*, Vol. 32 No. 3, pp. 163-71.

White, H.D. and McCain, K.W. (1997), "Visualization of literatures", in Williams, M.E. (Ed.), *Annual Review of Information Science and Technology*, Vol. 34, pp. 99-168.

White, H.D. and McCain, K.W. (1998), "Visualizing a discipline. An author co-citation analysis of information science, 1972-1995", *Journal of the American Society for Information Science*, Vol. 49 No. 4, pp. 327-55.

White, M.D. and Wang, P. (1997), "A qualitative study of citing behaviour: contributions, criteria, and metalevel documentation concerns", *Library Quarterly*, Vol. 67 No. 2, pp. 122-54.

Wilson, C.S. (1999), "Informetrics", in Williams, M.E. (Ed.), *Annual Review of Information Science and Technology*, Vol. 34, pp. 107-247.

Woods, W.A. (1997), "Conceptual indexing: a better way to organize knowledge", *Sun Labs Technical Report: TR-97-61*, Sun Microsystems Laboratories, Mountain View, CA.

Further reading

Bibexcel (2001), Bibexcel software: a tool-box developed by Olle Persson, Inforsk, Umeå university, Sweden, available at: www.umu.se/inforsk/Bibexcel/.

Frankfort-Nachmias, C. and Nachmias, D. (1997), *Research Methods in the Social Sciences*, 5th ed., Edward Arnold, London.

Glänzel, W. and Czerwon, H.-J. (1995), "A new methodological approach to bibliographic coupling and its application to research front and other core documents", in Koening, M.E.D. and Bookstein, A. (Eds), *Fifth Biennial Conference of the International Society for Scientometrics and Informetrics, River Forest, IL., USA, 1995,* 7-10 June, Learned Information, Medford, NJ, pp. 167-176.

Glänzel, W. and Czerwon, H.-J. (1996), "A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level", *Scientometrics*, Vol. 37 No. 2, pp. 195-222.

This article has been cited by:

1. Alex Fabianne de Paulo, Geciane Silveira Porto. 2017. Solar energy technologies and open innovation: A study based on bibliometric and social network analysis. *Energy Policy* **108**, 228-238. [Crossref]
2. Alex Fabianne de Paulo, Luísa Cagica Carvalho, Maria Teresa G.V. Costa, Jose Eduardo F. Lopes, Simone V.R. Galina. 2017. Mapping Open Innovation: A Bibliometric Review to Compare Developed and Emerging Countries. *Global Business Review* **18**:2, 291-307. [Crossref]
3. ur RehmanSajjad, Sajjad ur Rehman, AlajmiBibi, Bibi Alajmi. 2017. Knowledge organization content in graduate coursework. *Library Review* **66**:1/2, 90-106. [Abstract] [Full Text] [PDF]
4. Daniele Santoni, Elaheh Pourabbas. 2016. Automatic Detection of Words Associations in Texts Based on Joint Distribution of Words Occurrences. *Computational Intelligence* **32**:4, 535-560. [Crossref]
5. Isabela Neves Ferraz, Nathália de Melo Santos. 2016. The relationship between service innovation and performance: a bibliometric analysis and research agenda proposal. *RAI Revista de Administração e Inovação* **13**:4, 251-260. [Crossref]
6. Chris D. Paice. Lexical Analysis of Textual Data 1-6. [Crossref]
7. Dietmar Wolfram. 2015. The symbiotic relationship between information retrieval and informetrics. *Scientometrics* **102**:3, 2201-2214. [Crossref]
8. Dangzhi Zhao, Andreas Strotmann. 2015. Analysis and Visualization of Citation Networks. *Synthesis Lectures on Information Concepts, Retrieval, and Services* **7**:1, 1-207. [Crossref]
9. Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology* **65**:9, 1820-1833. [Crossref]
10. Rhoda C. Joseph. 2013. A structured analysis of e-government studies: Trends and opportunities. *Government Information Quarterly* **30**:4, 435-440. [Crossref]
11. Tomaz Bartol. 2012. Assessment of indexing trends with specific and general terms for herbal medicine. *Health Information & Libraries Journal* **29**:4, 285-295. [Crossref]
12. Jenny A. Glikman, Beatrice Frank. 2011. Human Dimensions of Wildlife in Europe: The Italian Way. *Human Dimensions of Wildlife* **16**:5, 368-377. [Crossref]
13. Jadhav Vandana Sheshrao, V. S. Khaparde. 2011. Citation Analysis of Ph.D. Theses on Physics Submitted to Dr. Babasaheb Ambedkar Marathwada University. *Collnet Journal of Scientometrics and Information Management* **5**:1, 115-127. [Crossref]
14. Barbara Schultz-Jones. 2009. Examining information behavior through social networks. *Journal of Documentation* **65**:4, 592-631. [Abstract] [Full Text] [PDF]
15. Omwoyo Bosire Onyancha, Dennis N. Ocholla. 2009. Is HIV/AIDS in Africa distinct? What can we learn from an analysis of the literature?. *Scientometrics* **79**:2, 277-296. [Crossref]
16. Kristie Saumure, Ali Shiri. 2008. Knowledge organization trends in library and information studies: a preliminary comparison of the pre- and post-web eras. *Journal of Information Science* **34**:5, 651-666. [Crossref]
17. Katherine W. McCain, Laura J. Salvucci. 2006. How influential is Brooks' Law? A longitudinal citation context analysis of Frederick Brooks' The Mythical Man-Month. *Journal of Information Science* **32**:3, 277-295. [Crossref]

18. Fidelia Ibekwe-SanJuan. 2006. Constructing and maintaining knowledge organization tools: a symbolic approach. *Journal of Documentation* **62**:2, 229-250. [Abstract] [Full Text] [PDF]